CS5500

Reinforcement Learning (RL)

Assignment No. 1

Name - Raktim Gautam Goswami

Roll number - EE17BTECH11051

Q1)

Solⁿ - a) There are 3 states - Tall, short, medium.

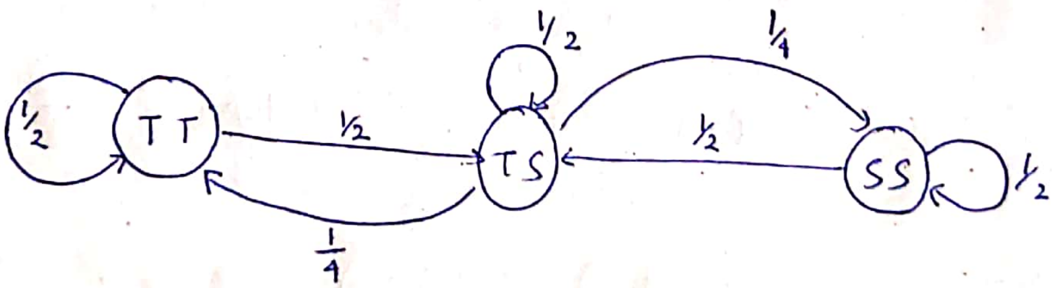
~~If indiv state~~

If individual is tall, height of offspring $= \begin{cases} \text{Tall} & \text{w.p. } \frac{1}{2} \\ \text{medium} & \text{w.p. } \frac{1}{2} \\ \text{short} & \text{w.p. } 0 \end{cases}$

If individual is medium, height of offspring $= \begin{cases} \text{Tall} & \text{w.p. } \frac{1}{4} \\ \text{medium} & \text{w.p. } \frac{1}{2} \\ \text{short} & \text{w.p. } \frac{1}{4} \end{cases}$

If individual is short, height of offspring $= \begin{cases} \text{Tall} & \text{w.p. } 0 \\ \text{medium} & \text{w.p. } \frac{1}{2} \\ \text{short} & \text{w.p. } \frac{1}{2} \end{cases}$

$\therefore$ the state-space diagram is as follows -



The transition probability matrix is as follows.

$$P = \begin{array}{c} \\ T \\ M \\ S \end{array} \begin{array}{c} \begin{array}{ccc} T & M & S \end{array} \\ \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \end{array}$$

b) ~~Probabilities that the offspring of~~

Required probability of first generation offspring

$= \text{second row of } P \text{ matrix} = \begin{cases} \frac{1}{4} & \text{for tall} \\ \frac{1}{2} & \text{for medium} \\ \frac{1}{4} & \text{for } \cancel{\text{small}} \text{ short.} \end{cases}$

Required probability of second generation offspring

$= \text{second row of } P^2 \text{ matrix} = \begin{cases} \frac{1}{4} & \text{for tall} \\ \frac{1}{2} & \text{for medium} \\ \frac{1}{4} & \text{for short} \end{cases}$

~~Requi~~ Similarily,

required probability of third generation offspring

$= \text{second row of } P^3 \text{ matrix} = \begin{cases} \frac{1}{4} & \text{for tall} \\ \frac{1}{2} & \text{for medium} \\ \frac{1}{4} & \text{for short} \end{cases}$

c) Clearly, there ~~follows~~ is a trend.

$\therefore$ Required probability $=$ second row of P^n matrix

$= \begin{cases} \frac{1}{4} & \text{for tall} \\ \frac{1}{2} & \text{for medium} \\ \frac{1}{4} & \text{for short.} \end{cases}$

Q2)

Solⁿ- a) The states are

S, 1, 3, 5, 6, 7, 8, W

The transition matrix (P) is as follows -

	S	1	3	5	6	7	8	W
S	0	1/6	1/6	1/6	1/6	1/6	1/6	0
1	0	0	1/6	1/6	1/6	2/6	1/6	0
3	0	0	1/6	1/6	1/6	1/6	2/6	0
5	0	0	1/6	1/6	1/6	1/6	1/6	1/6
6	0	0	1/6	0	2/6	1/6	1/6	1/6
7	0	0	1/6	0	0	3/6	1/6	1/6
8	0	0	1/6	0	0	0	4/6	1/6
W	0	0	0	0	0	0	0	1

b) The only absorbing state is W

c) We know, $V = R + \gamma P V$; [We have P from part (a)]

Let $\gamma = 0.9$

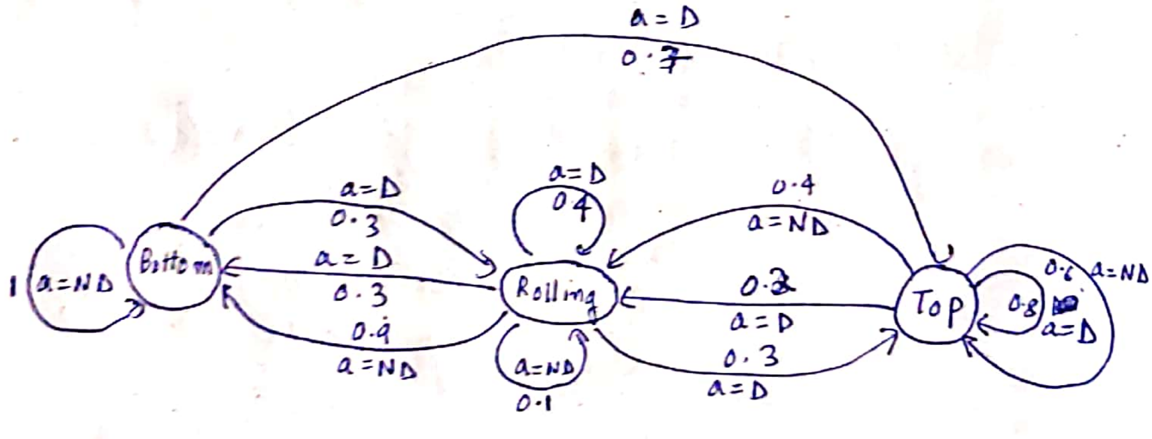
$R = [0\ 0\ 0\ 0\ 0\ 0\ 0\ 1]^T$, i.e., reward is 0 for all states except for state W

$\therefore V = (I - \gamma P)^{-1} R$

$= [4.84, 4.94, 4.94, 5.60, 5.60, 5.60, 5.60, 10]^T$

Q3)

Sol: a)



D → Driving
ND → Not driving

b) A deterministic policy is

$$\pi(s) = \begin{cases} \text{Drive w.p. 1} & \text{when } s = \text{Bottom} \\ \text{Drive w.p. 1} & \text{when } s = \text{Rolling} \\ \text{Dont drive w.p. 1} & \text{when } s = \text{Top} \end{cases}$$

A stochastic policy is

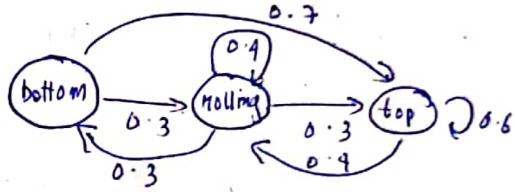
$$\pi(a|bottom) = \begin{cases} 0.5 & \text{for } a = \text{Driving} \\ 0.5 & \text{for } a = \text{Not driving} \end{cases}$$

$$\pi(a|rolling) = \begin{cases} 0.5 & \text{for } a = \text{Driving} \\ 0.5 & \text{for } a = \text{Not driving} \end{cases}$$

$$\pi(a|top) = \begin{cases} 0.5 & \text{for } a = \text{Driving} \\ 0.5 & \text{for } a = \text{Not driving} \end{cases}$$

c) For deterministic policy, transition probability matrix is

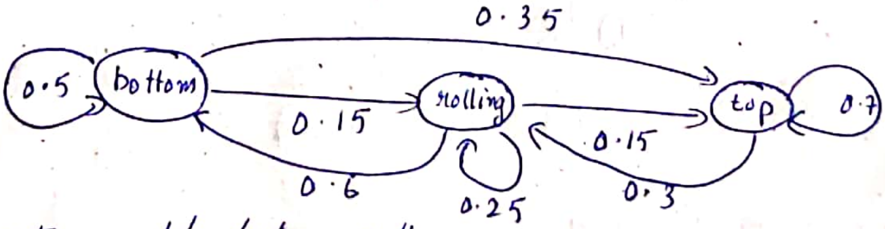
$$P = \begin{array}{c} \\ \text{bottom} \\ \text{rolling} \\ \text{top} \end{array} \begin{array}{c} \begin{array}{ccc} \text{bottom} & \text{rolling} & \text{top} \end{array} \\ \begin{bmatrix} 0 & 0.3 & 0.7 \\ 0.3 & 0.4 & 0.3 \\ 0 & 0.4 & 0.6 \end{bmatrix} \end{array}$$



For stochastic policy, each element of matrix is determined as follows -

$$P(s'|s) = \sum \left(\pi(a|s)\, P(s'|s,a)\right)$$

$$\therefore P = \begin{array}{c} \\ \text{Bottom} \\ \text{rolling} \\ \text{top} \end{array} \begin{array}{c} \begin{array}{ccc} \text{bottom} & \text{rolling} & \text{Top} \end{array} \\ \begin{bmatrix} 0.5 & 0.15 & 0.35 \\ 0.6 & 0.25 & 0.15 \\ 0 & 0.3 & 0.7 \end{bmatrix} \end{array}$$



d) A history dependent policy is

$\cancel{\pi(a|s_{t-1}, s_2, \ldots s}$

$$\pi(a|s_t, s_{t-1}, s_{t-2}, \ldots s_1) = \begin{cases} 0.3 & \text{when } r_t < \sum_{i=1}^{t-1} r_i \\ 0.4 & \text{when } r_t = \sum_{i=1}^{t-1} r_i \\ 0.3 & \text{when } r_t > \sum_{i=1}^{t-1} r_i \end{cases}$$

Q4)

$0.9 \times 0.9 + 0.6 \times 0.1$

$0.4 \times 0.1 + 0.6 \times 0$

Sol^n - For policy π_1,

$$P_{\pi_1} = \begin{array}{c|c} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{bmatrix} 0 & 0.9 & 0.1 & 0 \\ 0.1 & 0 & 0 & 0.9 \\ 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0.1 \end{bmatrix} \end{array}$$

For policy π_2

$$P_{\pi_2} = \begin{array}{c|c} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{bmatrix} 0 & 0.1 & 0.9 & 0 \\ 0.9 & 0 & 0 & 0.1 \\ 0.1 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

For policy π_3

$$P_{\pi_3} = \begin{array}{c|c} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{bmatrix} 0 & 0.42 & 0.58 & 0 \\ 0.1 & 0 & 0 & 0.9 \\ 0.1 & 0 & 0 & 0.9 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$R = [-10 \quad -10 \quad -10 \quad 100]^T$$

a) Assuming $\gamma = 0.9$,

$$V = (I - \gamma P)^{-1} R$$

$\therefore V^{\pi_1} = [755 \quad 867 \quad 691 \quad 1000]^T$

$V^{\pi_2} = [755 \quad 691 \quad 867 \quad 1000]^T$

$V^{\pi_3} = [772 \quad 869 \quad 869 \quad 1000]^T$

b) π_3 seems to be the best one as here all the values of V^{π_3} are higher compared to others

c) π_2 & π_1 are not ~~[illegible]~~ comparable as here in V^{π_2} & V^{π_3} two ~~[illegible]~~ states are having same values & in other 2 states, the values are interchanged. However, π_3 can be compared as V^{π_3} has all elements greater than V^{π_2} & V^{π_1} $\therefore$ it is the best policy.

Q 5)

Sol^n- a) We can do this using either value iteration or policy iteration. In this case, we will solve using value iteration.

Step I:

We initialise $V_1(s)$ for $s \in S$, a small value ϵ.

Step II:

For all states, we find

$$V_{k+1}(s) \longleftarrow \max_a \left[\sum_{s' \in S} P^a_{ss'} \left(R^a_{ss'} + \gamma V_k(s') \right) \right]$$

Step III:

If $|V_{k+1}(s) - V_k(s)| < \epsilon$ for all $s \in S$,

we go to step IV, (set $V_{k*}(s) = V_*(s)$)

else,

we go back to step II.

Step IV:

For all states $s \in S$,

$$\pi_*(s) = \arg\max_a V_*(s)$$

$\therefore$ we get $\pi_*(s)$ such that $V^{\pi_*}(s)$ is maximum for all state of the MDP.

Q6) The python code for this question is attached with the submission

Sol^n - a) For $\gamma = 1$, we cannot find an optimal value or policy as it becomes an infinite process and value function becomes unbounded.

b) For $\gamma = 0.9$,

$$V_* = [65.6, 72.8, 80.9, 89.9, 99.9, 99.9]^T$$

Using $\pi^* = \max_a V_*^{\circ}$

$\pi^*(s_1) = \begin{cases} 1 & \text{for } \cancel{\text{right}} \text{ moving right} \\ 0 & \text{" " left} \end{cases}$

$\pi^*(s_2) = \pi^*(s_3) = \pi^*(s_4) = \pi^*(s_5) = \begin{cases} 1 & \text{for moving right} \\ 0 & \text{" " left} \end{cases}$

$\pi^*(s_6) = 1$ for staying at s_6

For $\gamma = 0.5$,

$V_* = [1.24 \quad 2.49 \quad 4.99 \quad 9.99 \quad 19.99 \quad 19.99]^T$

$\pi^*(s)$ is same as for $\gamma = 0.9$

For $\gamma = 0.1$,

$V_* = [0.0011 \quad 0.011 \quad 0.11 \quad 1.1 \quad 11 \quad 11]^T$

$\pi^*(s)$ remains same as for $\gamma = 0.9$.

$\therefore$ it can be observed that the optimal value function changes for different γ. However, policy remains the same.

c) Adding c to all rewards is ~~to the~~ found out by ~~the~~ running the program in python.

For $c = -1$, $\gamma = 0.9$,

$V_* = [55.6 \quad 62.8 \quad 70.9 \quad 79.9 \quad 89.9 \quad 89.9]^T$

For $c = 1$, $\gamma = 0.9$,

$V_* = [75.6 \quad 82.8 \quad 90.9 \quad 99.9 \quad 109.9 \quad 109.9]^T$

For $c = 10$, $\gamma = 0.9$

~~$V_* = [75.6 \quad 82.8 \quad 90.9 \quad 99.9 \quad 109.9 \quad 109$~~

$V_* = [165.6 \quad 172.8 \quad 180.9 \quad 189.9 \quad 199.9 \quad 199.9]^T$

$\because$ the same value ~~reward~~ is being added to all the rewards, the policy remains the same.

d) For any policy π,

$V^\pi = (I - \gamma P)^{-1} R$

$$\hat{V}^\pi = (I - \gamma P)^{-1}(R + C) \quad ; \text{ where } C = [\underbrace{c \; c \; \cdots \; c}_{\text{No. of states}}]^T$$

$\therefore$ clearly, $\hat{V}^\pi = V^\pi + (I - \gamma P)^{-1} C$

Q8)

Sol^n - $L(V) = \max_{a \in A} [R^a + \gamma P^a V]$

$\because V_*$ is a fixed point of operator L,

$\therefore T(V_*) = V_*$

$$|V_{k+1} - V_*|_\infty = |T(V_k) - T(V_*)|_\infty$$

$$= \left|\max_{a \in A} [R^a + \gamma P^a V_k] - \max_{a \in A} [R^a + \gamma P^a V_*]\right|_\infty$$

$$\leq \max_{a \in A} \left|[R^a + \gamma P^a V_k] - [R^a + \gamma P^a V_*]\right|_\infty$$

$$= \gamma |P^a (V_k - V_*)|_\infty$$

$$\leq \gamma |V_k - V_*|_\infty$$

$$\therefore |V_{k+1} - V_*|_\infty \leq \gamma |V_k - V_*|_\infty$$

$$\leq \gamma^2 |V_{k-1} - V_*|_\infty$$

$$\leq \gamma^3 |V_{k-2} - V_*|_\infty$$

$$\leq \gamma^k |V_1 - V_*|_\infty$$

$$\therefore |V_{k+1} - V_*|_\infty \leq \gamma^k |V_1 - V_*|_\infty$$

Scanned by CamScanner

Q7)

Sol^n – The value for any state can be found out by ~~using~~ recursively running

$$V^\pi(s) \leftarrow \max_{a \in A} \left\{ \sum_{s' \in S} P^a_{ss'} \left[R^a_{ss'} + \gamma V^\pi(s') \right] \right\}$$

This is carried on for ~~all~~ all the states with different values of γ and P. ($\because$ on changing noise, P changes)

$\therefore$ The required parameters are as follows –

For close exit, risking the cliff, $\gamma = 0.1$, noise $= 0.45$

For distant exit, risking the cliff, $\gamma = 0.99$, noise $= 0.5$

For close exit, avoiding the cliff, $\gamma = 0.99$, noise $\simeq 0$

For ~~cl~~ distant exit, avoiding the cliff, $\gamma = 0.1$, noise $= 0$

Scanned by CamScanner